Shiv Chandaria
Assignment 2
Word Count : 1040

DA 201 – Covid Vaccines Analysis

The UK government wants to implement a marketing campaign to promote the second dose of the COVID vaccines. To plan the campaign successfully, the government wants to understand what the total 1st and 2nd vaccinations are across time in different regions, the recovery and death rates, trending topics on twitter, and regions which have not yet reached a peak in hospitalisations. A successful campaign will increase the amount of people who will have taken both vaccines as it will help the government's goal of ensuring public health.

The government has contextualised the problem clearly. With the goal of not undermining the campaign, I would seek to find out from the government how the data was collected, validate its sources for accuracy, and ensure that all the data does not breach any rules of privacy.

Analytical Approach

Before working with the data, I familiarized myself with the metadata.

After importing all the necessary Python dictionaries into my Jupyter notebook, I imported the 'covid_19_u_cases.csv', and the 'covid_19_uk_vaccinated.csv' as Panda dataframes. I began exploring these dataframes through looking at the shape, values and datatypes and noted that both tables had the same number of rows, and identical first 8 columns. This indicated that I could outer join the two dataframes to create one dataframe which would help with my future analysis.

Once I had the created the joined dataframe, I began the data wrangling process. I began by deleting unnecessary columns. When looking at the data types I noticed that the date column was an object. I therefore changed the data type to a datetime format. When looking for missing values, I noticed that there were only 8 missing values located in two rows. I proceeded to find these rows, and after observing the missing data, and the data around, I determined that using a forward fill would be the most appropriate method to deal with the missing data. As I continued my exploratory analysis, I wanted to understand the data in the 'Deaths, Cases, Recovered, Hospitalised, Vaccinated, First Dose and Second Dose' columns. I created a smaller dataframe with just information from Gibraltar, and after looking at the table in its entirety I understood how the data changed overtime. I made the following observations:

- Vaccine program began on 11th Jan 2021
- Data for Deaths, Cases and Recovered seems to be cumulative (a sum of total incidents till that date)
- Data for Hospitalised, Vaccinated, First Dose, and Second Dose is daily total
- Data for recoveries suddenly stops on 4th August, 2021

As the government wanted to understand the difference between first dose and second dose uptake in different regions, I created another dataframe which grouped data by 'Province/State' and summed the first dose and second dose taken in each region. I also added two more columns which calculated the difference between total first dose and second dose, and the percentage of people who have not taken a second dose. Corresponding graphics were created.

To understand deaths and recoveries, I created visuals to reveal how they have changed over time. For recoveries, I limited the timeframe till August 2021 as that is when data stopped being collected. For deaths, I noticed that data from the 'others' region was skewing the data, so I excluded it from my final graphic.

To understand trending topics on twitter, I used the 'tweets.csv' data provided by the government. After importing and exploring the data, I noted that the tweet content and hashtags were under the 'text' column. I wrote a function which would pull hashtags from the text column and store them in a series. After writing code which counted the number of instances each hashtag appeared, a dataframe showing all hashtags which were used over 100 times was created. A bar graph showing these hashtags and trending topics was then produced.

Lastly, I continued the work of the previous consultant to understand if there was to be another peak in hospitalisations. At the government's request I focused on the Channel Islands, and created a graphic based on the consultant's code. The graphics shows a 30-day moving average of hospitalisations in the Channel Islands. Following the trending average, one can see that hospitalisations are on the rise, and may lead to another peak.

Visualisations, Insights, Patterns and Trends

Several visualisations were created to tell the complete story that the data highlights and respond to the government's questions. Two bar graphs were created showing the amount of people partially vaccinated, and the percentage of people partially vaccinated by region. The colours used for each region are the same in both bars. The data reveals that while there is variance in the amount of people who are partially vaccinated in each region, the percentage of partially vaccinated people in all regions is around 4.5%. The region with the highest number of partially vaccinated people is Gibraltar.

Two line charts, one with data from all regions, and the second with data from Gibraltar, were also created to see how first dose and second dose vaccinations changed over time. When observing the Gibraltar data, there was a peak in February 2021 for the first dose, and a peak in April and May 2021 for the second dose. This is similar to the data from all regions.

Two more line charts were created showing deaths and recoveries over time. The recovery line chart (which ends in August), shows that Montserrat, Cayman Islands, and Saint Helena, have had very few recoveries. The death chart reveals that there are certain periods of time when the death rate accelerates; like from March 2020 to May 2020, and December 2020 to February 2021.

To visualise the twitter data, a bar chart showing the instances of most commonly used hashtags is created. The bar graph is ordered in descending order to highlight the most popular topics which happen to be all related to the Covid-19 other than one for Greece and another for China.

The last visual is a line chart showing the rolling 30 day average of hospitalisations in the Channel Islands. Based on the visual it seems unlikely that hospitalisations will reach another peak, but the gradual increase is something to keep an eye on.